Methodological and Statistical Developments for EMA Databases

Theodore A. Walls, Ph.D.

The Methodology Center

The Pennsylvania State University

Acknowledgements

- This work was funded by a NIDA center grant for methodological research on intensive measurement databases (Grant # P50 DA10075).
- Portions of this work are synopses of collaborative work of the Intensive Longitudinal Data Study Group at Penn State

  - Richard Li, PSU, Department of Statistics
  - Steve Rathbun, PSU, Department of Statistics
  - Wayne Osgood, PSU, Department of Sociology
  - Tammy Root, PSU, The Methodology Center
  - Steve Boker, Notre Dame, Department of Quantitative Psychology
  - Joseph Schafer, PSU, Department of Statistics
  - Linda Collins, PSU, The Methodology Center
  - Ted Walls, PSU, The Methodology Center, Study Group Organizer

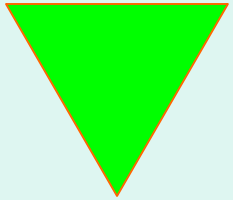
  - At-large: Saul Shiffman, University of Pittsburg; Richard Jones, University of Colorado; Peter Hovmand, Washington University

Outline

- Part 1: The need for models for intensive longitudinal data
- Part 2: Ways to estimate series more precisely
  - Loess
  - FDA
- Part 3: Some New Core Techniques
  - Dynamical Systems Models
  - Multi-level Time-Series models
  - Control Systems Techniques
  - Point Process Models
- Part 4: Open Questions and Future Possibilities

Part 1:

The Need for Models for Intensive Longitudinal Data



Types of Scientific Questions

- What is a person's average level of Y? Does this level differ over time among persons? Can we explain these differences in levels? [ANOVA}
- What is a person's rate of change on Y? Does this rate differ over time among persons? Can we explain these differences in rates? [Growth Models]
- Can we explain patterns of change in Y from multiple X's, for different parts of the series, for different subjects, controlling for previous levels of Y? [Mixed, multi-level models]

Based in part on Bolger, Davis & Rafaeli, 2003
and Schwartz& Stone, 1998

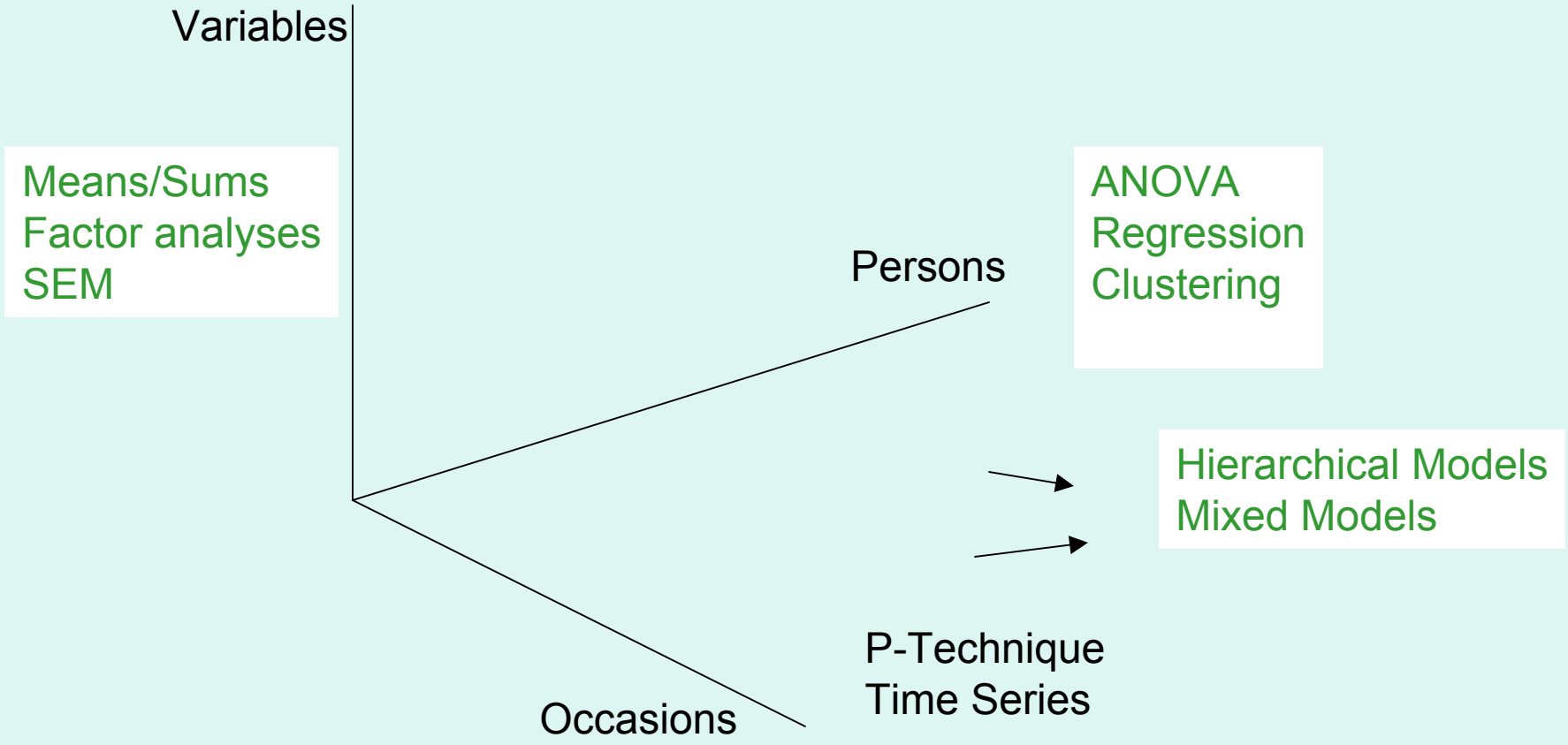
Current Statistical Approaches

- For design that looks like:

  O O O O O O O O O maybe X O O O O …

  Maybe with more than one group.
  - Using analyses that fit lines through many (within person) occasions of measurement, with separate mean and variance components computed for between person differences
  - Concluding that x predicts for y over many occasions, where y may be a rate or an adjusted rate, maybe have some covariates
  - Discuss variance attributable to within and between person sources as understood by multi-level models

Multiple Dimensions of EMA Data: Cattell's Data Box



Source: Cattell, 1966

Statistical Challenges in EMA Data

- First, currently available statistical techniques are not well-suited to the analysis of data reflecting more than 15-20 occasions.
- Second, because these data collection devices are “smart”, both the frequency and spacing of measurements (e.g. measured at all occasions or not; even or uneven spacing, etc.), and the nature of the underlying distributions of the measurement variables (particularly, Gaussian or non-Gaussian) may vary enormously.
- Third, with the emergence of these databases readily arise questions of how to conceptualize and aggregate patterns of change across subjects, groups, and clusters.

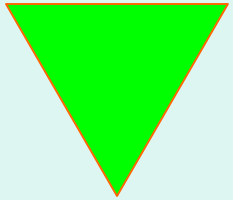
Statistical Challenges in EMA Data

- Fourth, most databases of this type include markers of time-varying and time invariant covariates.
- There may be new ways to model the intensity of events (e.g., as reflected in the arrival of recurrent or nonrecurrent events) with hazard functions and/or mean rates of occurrence over time.
- Also, new variable reduction techniques that characterize intra-individual variability are needed in light of the fact that data is available from so many occasions.

Part 2:

Ways to estimate series more precisely

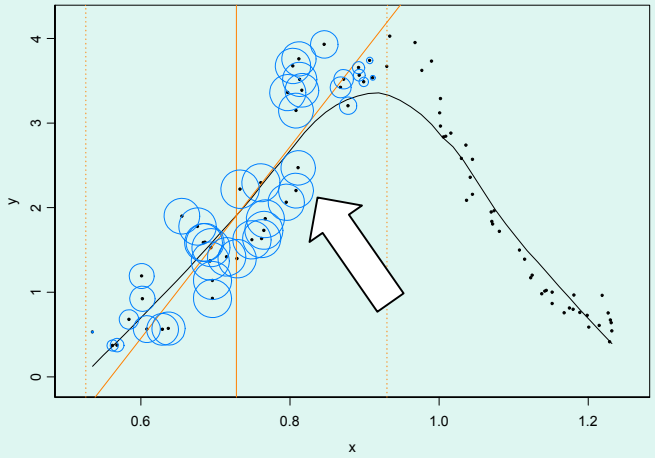
- Local Weighted Polynomial Regression (LOESS)
- Functional Data Analysis

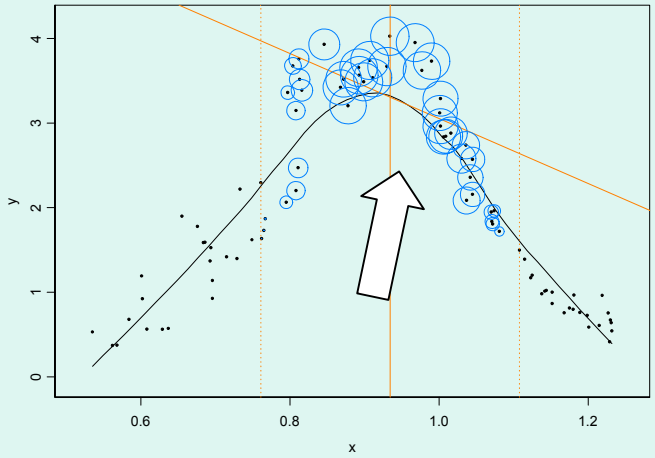


Local Weighted Polynomial Regression (LOESS)

- An advanced implementation of the general linear model
- Uses nearest neighbor strategy
  - A low order polynomial is fit to a subset of the data focusing on individual points for which a response is being estimated
  - These subsets are selected be setting the bandwidth to a given length (requires judgment based on data inspection and theory)
  - Programs (such as SAS) iterate through all points in the database to create a “smooth”
- Can think of this as taking advantage of some non-linear tricks, but still remaining in a linear (least squares) modeling framework

Visualizing LOESS





- The arrow shows the X value for which the estimate is being generated.
- The size of the circles show the amount of weight given to each score. Other values get 0 weight.
- The diagonal line is the weighted local regression function\

Source: Greg Snow, BYU, Department of Statistics
http://statweb.calpoly.edu/lund/stat430/classnotes/loessdemo.html

Advantages of LOESS

- The analyst does not need to specify the function for the regression equation (square, polynomials, etc.).
- For time series data, this is a great advantage because many of the curves of the series are difficult to fit
- Makes it possible to model complex processes well (based on the data) when a theory about the change does not exist
- Most procedures used with least squares modeling (multi-level versions, fit statistics, etc.) are retained.
- Easy to implement in most major statistical software packages; some nonparametric approaches are available

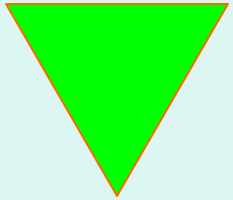
Functional Data Analysis

- Extension of LOESS in the sense that smoothing is required to get good enough estimates of curves
- Ramsay & Silverman, 1997, 2002.
  - Quantitative Psychologist at McGill and mathematician at Oxford
  - Assumption that human development is reflected in curves that can be expressed as functions; when we sample, we are sampling from those curves
  - Can evaluate functions that may influence each other
  - Can consider functions that interact, for example, knee and hip function
  - Possible to consider curves from multiple datasets simultaneously.

Part 3:

The Need for Models for Intensive Longitudinal Data

- Dynamical Systems Models
- Multi-level Time-Series models
- Control Systems Techniques
- Point Process Models



Dynamical Systems Models

- Interest in intrainvidual regulatory processes and how they change.
- Using structural equation models to analyze covariance among the value of a change score, its rate of change (1st derivative) and acceleration in the change (2nd derivative); Assess the affect of these state variables on another, an “attractor “
- Boker (1998) has estimated models for two variables, called a coupled oscillator
- Similar models developed by McArdle & Hamagami (1991)

An example of two coupled oscillators

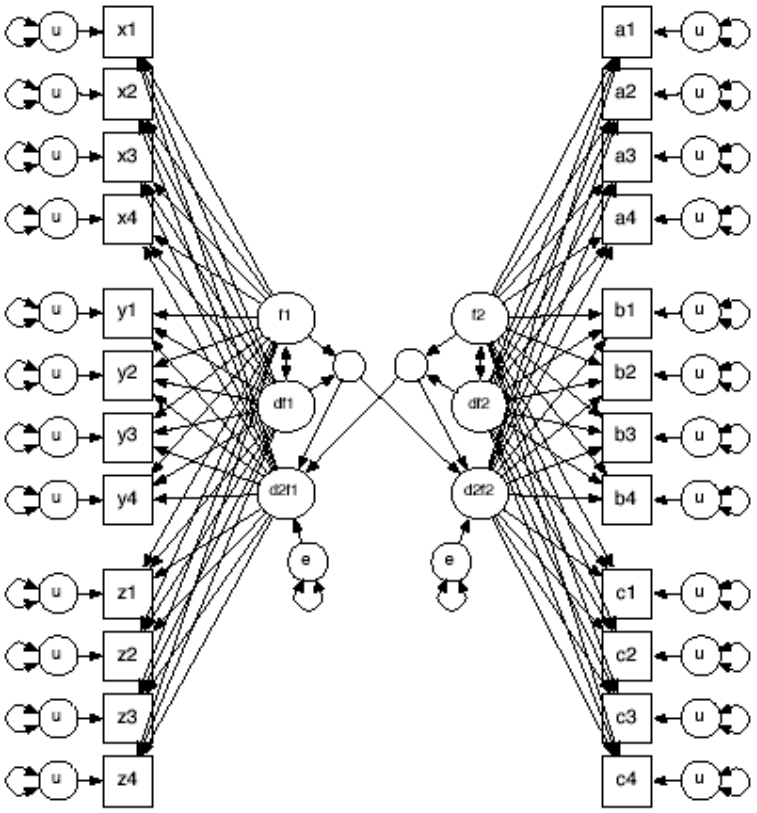


Figure 8. Path diagram of a multivariate latent differential equations model with two coupled latent variables.

Multi-level Time Series Models

- Utilizes lagging techniques
- Frequently utilize Moving Averages (the MA in ARIMA, ARMA)
- General strategy is to find the best time series AR(1), AR(2), etc.
- Interested in understanding of dynamic relationships—one series may lead to another; common approaches “state space modeling” and related ways to estimate these models (e.g. Kalman Filters)

Based on work by Rovine & Walls

Time Series Models for Multiple Series

- Advantages of Time Series
  - Conceived of for longer series
  - Information contained in the historical data of another series may allow understanding of dynamic processes if they are modeled jointly
  - Multi-level versions may allow interpretation of individual differences in these dynamic processes
  - Multivariate versions are increasingly possible

Based on work by Rovine & Walls

Control Systems Techniques

- Kalman Filters (sophisticated way of lagging; used in engineering routinely, helped with Saturn project to put man on the moon)
- Multiple variables in dynamic relationships
- Using differential equations to model time varying data
- Example: Lupus—symptom as the input and flare as the output. The flare causes change in treatment

Based on Work by J. Ramsay and D. Rivera

Control Systems Opportunities

- Adaptive Interventions
- Measured variable responses to intervention
- Differentiating dosage
  - Disturbance variables
  - Tailoring variables
  - Variable magnitudes

Temporal Point Process Models

- A Spatial Point Process is some kind of (random) process that generates locations in space at which something is found, e.g. a tree, a bird's nest, a measurement in pond water, etc.
- A Temporal Point Process is some kind of (random) process that generates events that occur over a period time.
- Because of the directional nature of time (there is a distinct past), there are a wider variety of models available for temporal or spatiotemporal point processes, than spatial only processes. The Poisson process, however, is mathematically equivalent regardless of whether you are working in space or time.

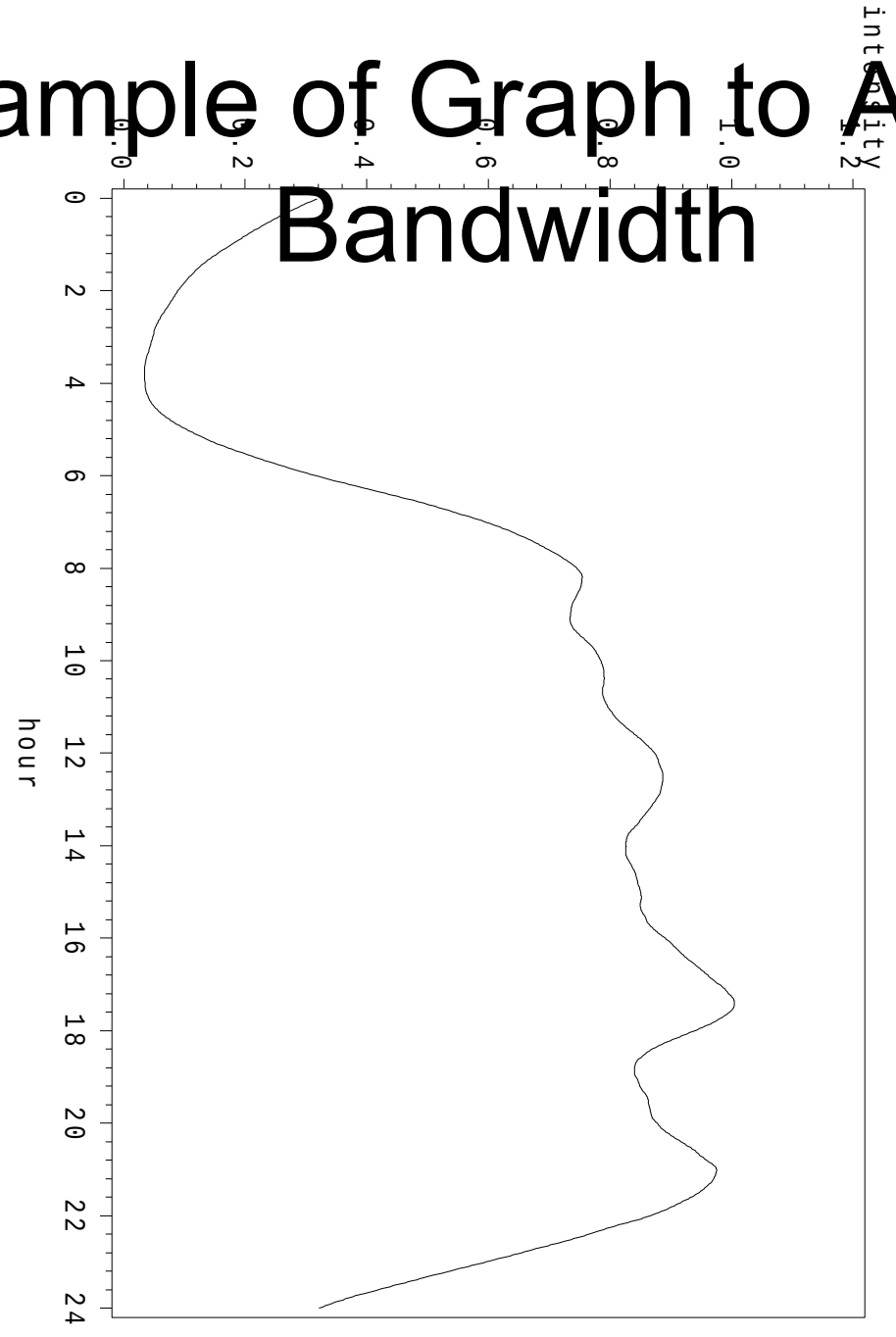
Temporal Point Process Models

- These models share some common conceptual ground with survival models
  - Survival models characterize the length of time to censoring for each subject and fit curves across subjects
  - Temporal process models also deal with the period of time and event occurrence.
- However, for a survival curve of many events, you would have to calculate the survival rate amidst many recurrent events. This makes it difficult to utilize all of the available information with these models.
  - The advantage of Temporal Point Process Models is that they utilize an “intensity function” that describes the recurrent events as an average number of events per unit time rather than multiple times to events.

Temporal Point Process Models

- How is a temporal point process defined?
  - Uses a “counting measure” N(A) –the number of events (A) occurring in an interval of time (N) e.g. per day.
  - The temporal point process is the joint distribution of all (disjoint) sets $N(A_k)$.
- How is the process estimated?
  - To define the intensity function, consider the expected number of events in a time interval divided by the length of that time interval. The intensity is defined by taking the limit as the length of that interval converges to zero. Thus, the intensity can be thought of as an "instantaneous rate of event occurrence.“
  - Modeling utilizes a Poisson process to describe the area under the curve. In the simplest model, independence of events is assumed; MLE is used for the estimation procedure.

Example of Graph to Assess Bandwidth



Temporal Point Process Models

- Advantages of this model
  - Efficiently characterizes continuous time information with respect to the intensity of the events over various increments of time (that the modeler selects)
  - Allows time-varying covariates; these are modeled in the same way as the predictors
  - Extensions will be able to allow random effects for individuals and also change the independence of event assumption inherent to the Poisson process.
  - May be able to build in simultaneous modeling of multiple levels of time (day, week, etc.)

Example Final Output Using Temporal Point Process Estimation

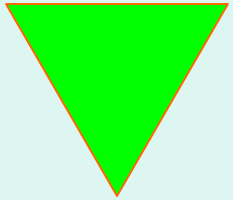
Results: The following table gives the estimates of the parameters for a modulated Poisson process, predicting the intensity of smoking events from Negative Affect, Arousal, and Attention.

Parameter	Estimate	SE	*z*
Intercept	1.43739	0.00807	
Negative Affect	0.01218	0.00850	1.43
Arousal	-0.02303	0.00851	2.71*
Attention	0.00559	0.00890	0.63

Data Source: S. Rathbun using Shiffman Data

Part 4:

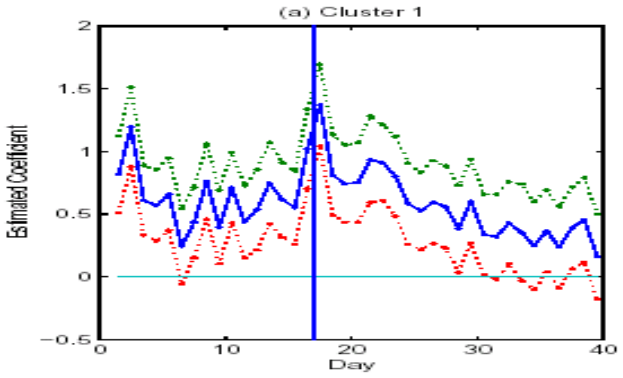
Open Questions and Future Possibilities

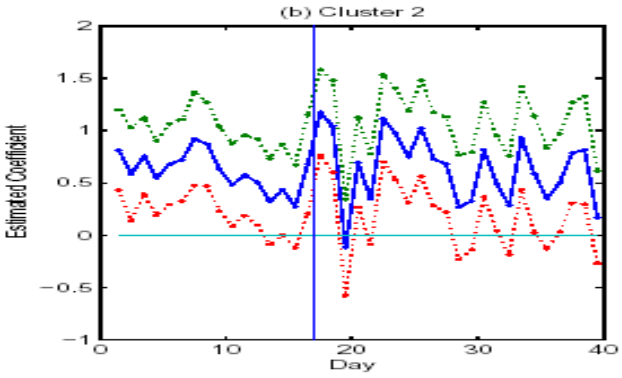


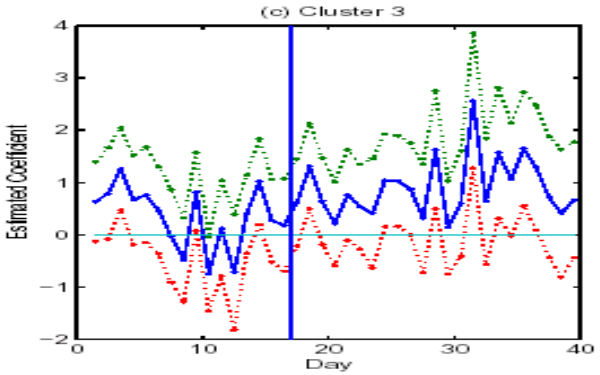
Open Questions

- Measurement—Classic issue of change, but worse because of so many occasions.
- When sampling occurs with diverse sources—may require multiple measurement models?
- How to set sampling density in order to detect something like do urges precede lapses, in anticipation of using a given technique.
- Power analyses for Subject, Time, & Variables
- How to handle reciprocal relations versus causality?
- The Heterogeneity vs. homogeneity trade-off

Example of Four different Clusters







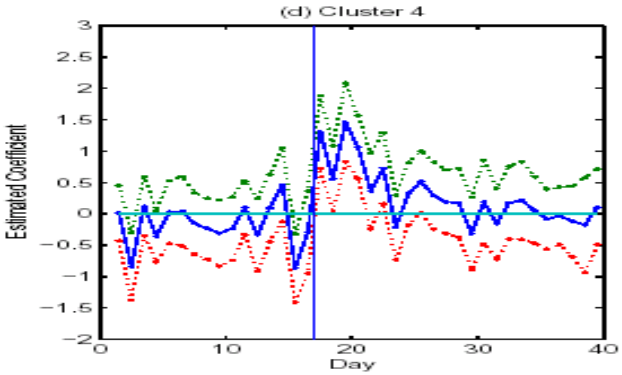


Figure 7: Plot of Estimated Effects of Negative Affect. Solid curves stand for estimate and dotted curves are 95% pointwise confidence intervals.

Source: Richard Li, From Shiffman data

Interrupted time-series

- Some questions with specific targeted analyses may not need many occasions of measurement…
  - Experimental ITS design; very robust design; handle maturation and X
  - If a slope change is the outcome and group comparison is the objective, then perhaps a mixed model can be run without excessive numbers of occasions

“Process Interruption”: A Simulation